DR. QUANJUN HU (Orcid ID : 0000-0001-6922-2144)

PROF. JIANQUAN LIU (Orcid ID : 0000-0003-4277-9772)

Received Date : 30-Nov-2016

Revised Date : 02-Mar-2017

Accepted Date : 02-Mar-2017

Article type : Original Article

# Chasing ghosts: Allopolyploid origin of *Oxyria sinensis* (Polygonaceae) from its only diploid congener and an unknown ancestor

Xin Luo[1†], Quanjun Hu[1, 2†], Pingping Zhou[1], Dan Zhang[1], Qian Wang[1,4], Richard J. Abbott[3], Jianquan Liu[1, 2*]

[1]MOE Key Laboratory for Bio-resources and Eco-environment, College of Life Science, Sichuan University, Chengdu, China.

[2]State Key Laboratory of Grassland Agro-Ecosystem, College of Life Science, Lanzhou University, Lanzhou, China.

[3]School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH, UK.

[4]Research Center for Medicine and Biology, Zunyi Medical University, Zunyi, China.

[†]Both authors contributed equally to this work

* Author for Correspondence: Jianquan Liu, Sichuan University, Chengdu, China. Phone: +86 18011455756. E-mail: liujq@nwipb.ac.cn.

**Abstract**

Reconstructing the origin of a polyploid species is particularly challenging when an ancestor has become extinct. Under such circumstances the extinct donor of a genome found in the polyploid may be treated as a 'ghost' species in that its prior existence is recognised through the presence of its genome in the polyploid. In this study, we aimed to determine the polyploid origin of *Oxyria sinensis* (2*n*=40) for which only one congeneric species is known, i.e. diploid *O. digyna* (2*n*=14). Genomic *in situ* hybridization (GISH), transcriptome, phylogenetic and demographic analyses, and ecological niche modeling were conducted for this purpose. GISH revealed that *O. sinensis* comprised 14 chromosomes from *O. digyna* and 26 chromosomes from an unknown ancestor. Transcriptome analysis indicated that following divergence from *O. digyna*, involving genome duplication around 12 million years ago (Ma), a second genome duplication occurred approximately 6 Ma to give rise to *O. sinensis*. *Oxyria sinensis* was shown to contain homologous gene sequences divergent from those present in *O. digyna* in addition to a set that clustered with those in *O. digyna*. Coalescent simulations indicated that *O. sinensis* expanded its distribution approximately 6-7 Ma, possibly following the second polyploidization event, whereas *O. digyna* expanded its range much later. It was also indicated that the distributions of both species contracted and re-expanded during the Pleistocene climatic oscillations. Ecological niche modeling similarly suggested that both species experienced changes in their distributional ranges in response to Quaternary climatic changes. The extinction of the unknown 'ghost' tetraploid species implicated in the origin of *O. sinensis* could have resulted from superior adaptation of *O. sinensis* to repeated climatic changes in the region where it now occurs.

**Key words**: *Oxyria*, allopolyploid speciation, ghost species, GISH, transcriptome, demographic history

## Introduction

Polyploidy, or whole-genome duplication (WGD), has played a key role in plant evolution and diversification (Grant 1981; Stebbins 1985; Otto 2007; Soltis *et al.* 2014; Soltis & Soltis 2016). All examined diploid angiosperms show genome duplication in their ancestry (Fawcett *et al.* 2009), while more than 15% of extant angiosperm species are considered to have originated directly via polyploidy, based on the lowest chromosome numbers reported for congeners (Wood *et al.* 2009). Polyploids can originate following hybridization between two different lineages (allopolyploidy) or via genome doubling of a single lineage (autopolyploidy) (Stebbins 1971; Lewis 1980; Grant 1981). Though the rate of autopolyploid formation is thought to be much higher than that for allopolyploid formation (Ramsey & Schemske 1998; Barker *et al.* 2016), a recent assessment of the relative frequencies of autopolyploids and allopolyploids among vascular plants indicates near parity of the two types (Barker *et al.* 2016), while in some genera allopolyploids are more frequent (Doyle & Sherman-Broyles 2016). This suggests that allopolyploids may hold an evolutionary advantage over autopolyploids. The establishment of a polyploid is aided by instant niche divergence from its progenitors (Fowler & Levin 1984; Ramsey 2011; Parisod & Broennimann 2016), which may originate more easily in allopolyploids than autopolyploids (Barker *et al.* 2016). Moreover,

allopolyploids may tolerate a wider range of environments and adapt to environmental change more readily due to their possession of two or more divergent genomes inherited from their progenitors (Barker *et al.* 2016).

Resolving the ancestry of polyploid species is important to an understanding of how such species evolved following their origin. However, such resolution becomes particularly challenging when an ancestor of a polyploid is extinct. Under such circumstances the extinct donor of a genome (or part of a genome) found in the polyploid may be treated as a 'ghost' species in that its prior existence is recognised through the presence of its genome (or part of it) in the polyploid. If fossils are available from which DNA can be extracted it may be possible to identify an extinct ancestor directly through sequence analysis. However, this is not usually an option due to the unavailability of such fossils.

In the study reported here, we examined the origin of *Oxyria sinensis* (Polygonaceae), a perennial herbaceous species native to the East Himalaya (Fig. S1). This region with multiple high mountains and deep valleys created by uplifts of the Himalaya and Qinghai-Tibet Plateau since the middle Miocene contains numerous endemic plant species and is recognized as a biodiversity hotspot (Myers 2000; Mittermeier *et al.* 2004; Wilson 1992; Wu 1998). However, polyploids are not as widespread in the area (Liu *et al.* 2001; Liu 2004; Nie *et al.* 2005) as might be expected for a region subjected to climatic oscillation during the Quaternary and where, in parallel, frequent contractions and expansions of species ranges have occurred (Brochmann *et al.* 2004). *Oxyria sinensis* mainly occurs in dry and hot valleys at altitudes ranging from 1600 m to 3800 m a.s.l. (Meng *et al.* 2015) and rarely on the slopes of high mountains. It is trioecious, containing hermaphrodite, male and female plants, which in addition to reproducing sexually can form large clones on cliff faces and crevices through rhizome

formation (Liu *et al.* 2007; Zhao & Yang 2008). The three sex types were recently shown to differ in drought tolerance (Yang *et al.* 2014). The only congener to *O. sinensis* is *O. digyna* (Mountain sorrel), which is a hermaphroditic, herbaceous, perennial species widely distributed throughout the Arctic and the mountain ranges of North America and Eurasia in diverse habitats at elevations from sea level to 4500 m a.s.l. (Fig. S1). Although both species are distributed in the East Himalaya with overlapping ranges (Fig.S1), they never occur sympatrically at the same site (Meng *et al.* 2015, Wang *et al*. 2016). According to a recent phylogeographic analysis (Wang *et al.* 2016), *O. digyna* originated in the East Himalaya, migrated westward into Western Europe and eastward into Russia and North America, and became circumarctic in its distribution before the Pleistocene (Fig. S1). There is also phylogenetic evidence that *O. digyna* and *O. sinensis* diverged from each other ~12 million years ago (Ma) (Sun *et al.* 2012). We began the research reported here by conducting a cytological analysis which confirmed that *O. digyna* was diploid ($2n$=14) whereas *O. sinensis* was polyploid ($2n$=40). We then conducted genomic *in situ* hybridization (GISH), transcriptome, phylogenetic and demographic analyses, and also ecological niche modeling (ENM) to address the following questions: (i) How did *O. sinensis* originate, i.e. through autopolyploidy or allopolyploidy? (ii) What is the ancestry of *O. sinensis* in terms of its progenitors? (iii) How do the demographic histories of *O. sinensis* and *O. digyna* compare and what effect might polyploidization have had on the demographic history of *O. sinensis*? (iv) How did *O. sinensis* and *O. digyna* change their distributions in response to earlier and recent climatic changes?

## Material and methods

### Plant material

We collected fresh leaves and seed of *O. sinensis* and *O. digyna* for molecular and cytological analyses from throughout the distributions of both species in the East Himalaya. Nine populations of *O. sinensis* and 12 populations of *O. digyna* were sampled (Fig. S2, Table S1). Records of altitude, latitude, and longitude for each locality were taken using an Etrex GIS monitor (Garmin) (Table S1). For molecular analysis, fresh leaves were dried in the field, and stored in silica gel. For cytological and transcriptome analyses, seeds were germinated to produce plants grown to a mature stage in a common garden. We also collected fresh leaves of *Rheum tanguticum* grown at Lanzhou University, Gansu Province, China, for comparison in the transcriptome analysis.

### Cytological analysis and genomic *in situ* hybridization (GISH)

For cytological analysis, at least three individuals representing each of three populations (populations 1, 3 and 4, Table S1) of *O. sinensis* and three populations (populations 13, 16 and 19, Table S1) of *O. digyna* were examined. Cytological analyses were performed using root tips to determine chromosome numbers (Jewell & Islam-Faridi 1994). Seeds of each individual were germinated and roots were pretreated with colchicine (0.01% w/v) for 2–3 h before fixing and storing in 3:1 ethanol – acetic acid. Chromosomes at mitotic metaphase of at least three well-spread root tip squashes were counted for each individual using an OLYMPUS BX53 microscope.

To confirm that *O. sinensis* contains the genome of *O. digyna*, we performed GISH analyses as described in Leitch *et al.* (1994) with minor modifications. Total genomic DNA was isolated from young leaves of *O. digyna* by the CTAB method (Doyle & Doyle 1987) and labeled with digoxigenin-UTP (DIG- Nick Translation Mix, Roche Applied Science). Well-spread metaphase chromosomes of *O. sinensis* from root tip cells were used for slide denaturation and hybridization. Labeled DNA was detected with FITC-conjugated avidin (DIG- detecting Mix, Roche Applied Science) and chromosomes were counterstained with DAPI in Vectashield antifade solution. Slides were examined by a PerkinElmer UItraVIEW VOX 3D Live Cell Image System and photographs were captured using the Olympus IPP software package and embellished in Photoshop Adobe System.

**Transcriptome analysis**

Total RNA was extracted according to the CTAB protocol (Doyle & Doyle 1987) from leaves of one individual of each *Oxyria* species after germinating seeds collected from populations 1 and 13 (Fig. S2) and also leaves of one individual of *Rh. tanguticum*. We also downloaded transcriptome data for *Rh. nobile* (Wang *et al.* 2014). The integrity of RNA was assessed using an Agilent 2100 Bioanalyzer. 20 μg of total RNA from each individual was purified using polydT conjugated beads to extract polyA-tagged mRNA. This was subsequently cleaved into ~200 bp fragments by treatment with divalent cations at 75°C. First and second strand cDNA synthesis was carried out using reverse transcriptase (Invitrogen) with random hexamer primers, RNase H (Invitrogen) and DNA polymerase I (New England BioLabs). Sequencing was performed on an Illumina HiSeq 2000. After removing

adapter sequences, raw reads were filtered using Trimmomatic V 0.32 (Bolger *et al.* 2014) to generate clean reads for all subsequent analyses. Trinity V 2.1.1 (Grabherr *et al.* 2011) was used to assemble paired-end short reads of exactly 200 bp into contigs. A set of non-redundant, representative sequences were produced by CD-HIT V 4.6.1 (Huang *et al.* 2010) with a threshold value of 0.95; cds sequences and pep sequences in the open read frame were predicted by the Transdecoder program of the Trinityrnaseq V 2.0.6. (http://trinityrnaseq.github.io/).

Orthologous sequences in the *Oxyria* species were determined using Inparanoid V 4.1. (Sonnhammer *et al.* 2015), while paralogous sequences were identified by conducting all vs all blast (blastn) comparisons within each species. We employed the approach of Blanc *et al.* (2004) to detect interspecific divergence and genome duplications (Jiao *et al.* 2012), and a maximum likelihood method implemented in the CODEML program of the PAML package V 4.1 (Yang 2007) to estimate rates of synonymous (Ks) and non-synonymous substitution (Ka) between paralogous and orthologous sequence pairs aligned by Prank (Löytynoja & Goldman 2008). We used the 1:1 orthologous sequence pairs of the two species to calculate interspecific genetic divergence and time of divergence based on Ks. We then dated genome duplications in the polyploid *O. sinensis* based on Ks between pairs of paralogs. Paralogs of a polyploid genome were sorted in order of relative age of duplication estimated by Ks between pairs of parologs (Lynch & Conery 2000). As each whole genome duplication event is expected to produce a normal distribution of Ks values, normal mixture models and the values of corresponding peaks could be estimated for Ks values between paralogs using the Mclust V.5.0.2 (Fraley *et al.* 2012; Scott *et al.* 2016) as implemented in R. A mutation rate

of $1.2 \times 10^{-8}$ for nuclear sequences estimated from fossil calibration of *Oxyria* (Wang *et al.* 2016) was used for all time estimations in addition to the lowest and highest mutation rates recorded for plants, i.e. $1 \times 10^{-8}$ and $1.5 \times 10^{-8}$ (Holliday *et al.* 2016; Ossowski *et al.* 2010).

To obtain further evidence that would help distinguish between an autopolyploid or allopolyploid origin of *O. sinensis* we examined phylogenetic relationships between homologous sequences of *O. sinensis* and related taxa, which included *O. digyna* and two diploid species of the sister genus, *Rh. tanguticum* and *Rh. nobile* ($2n = 22$). *Oxyria sinensis* ($2n = 40$) is likely to be a hexaploid when compared to diploid *O. digyna* ($2n=14$). Thus, two to three paralogs at some loci might be retained in *O. sinensis* due to genome triplication and incomplete loss of replicated genes. We extracted homologous sequences using Multiparanoid (Alexeyenko *et al.* 2006) across taxa according to the ratio of 1:1:1:2 and 1:1:1:3 between *Rh. tanguticum*, *Rh. nobile*, *O. digyna* and *O. sinensis*. For each group of homologous sequences at each locus, we constructed gene trees using PhyML 3.1 (Guindon *et al.* 2010) based on the method of maximum likelihood with sequences of *Rh. nobile* considered as outgroup.

**Phylogenetic relationships and demographic history based on analyses of seven nuclear loci**

Fourteen individuals from nine populations of *O. sinensis* were used in phylogenetic and demographic analyses of variation surveyed across seven nuclear loci (designated here as 1007, 3375, 6500, 7783, 8719, 9342 and 10702). These loci were previously surveyed in a study of populations of *O. digyna* (Table S1) (Wang *et al.* 2016). DNA extractions and PCR amplifications, using

homologous primers, were conducted following the procedures described by Wang *et al.* (2016). Amplified products were ligated into pMD19-T Vector with a pMD19-T Vector Cloning Kit (TaKaRa, Dalian, China) and 20 positive clones for each locus per individual were sequenced. Sequences for each nuclear gene were aligned and checked by MEGA 6.0 (Tamura *et al.* 2013) and by eye. All newly obtained sequences have been deposited in GenBank (Accession Numbers KX999721-KX999862). We also used materials and sequence data of *O. digyna* reported by Wang *et al.* (2016), but selected only populations of *O. digyna* that occurred in Asia for phylogenetic and demographic analyses because those occurring in Europe, North America and the Arctic were demonstrated to have originated and migrated from Asia (Wang *et al.* 2016).

Phylogenies were constructed for all nuclear haplotypes found in *O. sinensis* and *O. digyna* (with *Rh. tanguticum* as outgroup) using MrBayes v.3.0 (Ronquist & lsenbeck 2003) and with the GTR+G model selected. Four chains were run at the same time to 6 million generations. Trees were sampled every 100 generations. The posterior probabilities indicating support values for each branch were estimated following a 25% "burn-in", so the first 2500 trees were discarded as 'burn-in' and excluded from all inferences made. The output file was displayed in Treeview 1.6.6 (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

To investigate the early demographic histories of both *O. sinensis* and *O. digyna* in Asia, we designed and examined four plausible scenarios of demographic change using approximate Bayesian computation (ABC) in DIYABC V.1 (Cornuet *et al.* 2010) based on the sequence data of the seven nuclear loci. We excluded from this analysis other *O. digyna* populations outside Asia because during and following colonization of these other regions, they might have been subject to mutation fixation

rates and climatic oscillations different to those of Asian populations (Wang *et al*. 2016). The four demographic scenarios examined were (1) continuous population expansion since the origin of the two species, (2) recent expansion, (3) expansion-shrinkage and (4) expansion-shrinkage-expansion. DIYABC allows identification of which demographic scenario best fits the data and parameters of interest. The distribution of each parameter is listed in Table S3. A reference table containing $4\times10^6$ simulated data sets (on average $10^6$ per scenario) was generated. We used the 'pre-evaluation scenario prior combination' option in DIYABC to check which scenario the combination of priors produced simulated data sets that matched best the observed data. We used a mutation rate of $1.2\times10^{-8}$ (after Wang *et al.* 2016) and selected the scenario with the highest posterior probability by performing direct estimate and logistic regression analyses. Model checking for the best demographic scenario was conducted by simulating 10,000 datasets under each studied model to check the ability of a given scenario to produce data sets similar to the real data set.

**Ecological niche modeling of *O. digyna* and *O. sinensis***

Ecological niche modeling (ENM) was conducted via maximum entropy using MaxEnt 3.3.4 with default settings (Phillips & Dudík 2008) to examine ecological divergence between the two *Oxyria* species and to indicate their distributions currently and during the Last Glacial Maximum (LGM) and Last Interglacial (LIG) periods. Data used to perform ENM for current distributions included information from our field investigations and specimens' data sets (www.gbif.org) (excluding redundant and unreliable records) combined with data for 20 ecological variables (altitude

and 19 bioclimatic variables) downloaded from the WORLDCLIM database. Seven of the environmental variables (bio1 Annual Mean Temperature, bio2 Mean Diurnal Range, bio3 Isothermality, bio8 Mean Temperature of Wettest, bio15 Precipitation Seasonality, bio18 Precipitation of Warmest Quarter and bio19 Precipitation of Coldest Quarter) which exhibited pairwise Pearson correlation coefficients $r<0.7$ were selected for final analyses. Outputs were set at 25% for testing and 75% for training in the model. The accuracy of the model was measured by the area under the curve (AUC) of the receiver operating characteristic (ROC) plot (Fielding & Bell 1997). A model prediction was considered useful when an AUC score was in the range of 0.7–0.9, and good when >0.9 (Swets 1988). We set the maximum number of iterations to 5,000 and the number of replicates to 20. The output format was set to be logistic. DIVA-GIS version 7.5 (Hijmans *et al*. 2005) was used to draw the range of suitable distributions. ENMTools 1.3 (Warren *et al.* 2010) was used to estimate niche overlap between the two species assessed by Schoener's D and Warren's I similarity statistics, with 100 pseudo-replicates. Values of both D and I may range from 0 (no niche overlap) to 1 (identical niches).

## Results

### GISH analysis of *O. sinensis*

Chromosome counts showed that all plants examined of *O. digyna* were diploid with $2n = 14$ while those of *O. sinensis* were polyploid with $2n = 40$ (Fig. 1a and Fig. 1b). Chromosomes of *O. sinensis* probed with labeled total genomic DNA of *O. digyna* at mitotic metaphase showed that GISH

occurred for 14 of the 40 chromosomes (stained green, Fig. 1c). This confirmed that *O. sinensis* contains the entire *O. digyna* genome. GISH did not occur for the remaining 26 chromosomes of *O. sinensis* (stained blue, Fig. 1c), showing that these are derived from a different donor or donors. GISH therefore provided strong evidence for an allopolyploid origin of *O. sinensis*. Interestingly, there was no evidence for GISH occurring in any part of the 26 chromosomes that stained blue, indicating a lack of recombination and translocation between the *O. digyna* and other ancestral genome(s) contained within *O. sinensis*.

**Transcriptome analysis**

Clean sequences comprising 1.57G and 1.74G of cDNA were obtained for *O. digyna* and *O. sinensis*, respectively. The *de novo* assembly produced 36125 and 50955 contigs with an N50 value of 1209 and 1098 for *O. digyna* and *O. sinensis*, respectively (Table S2). We identified 14520 pairs of orthologous sequences using Inparanoid V 4.1 and the reciprocal best blast hit approach. The e-value was set at 1E-5 to separate paralogous from orthologous sequences and we identified 6420 and 15196 paralogs for *O. digyna* and *O. sinensis*, respectively, by all vs all blast (blastn) comparison. We calculated the Ka and Ks values for fourfold degenerate sites extracted from the 14520 pairs of orthologous sequences. Interspecific divergence based on the Ks values for 1:1 orthologous pairs between *O. digyna* and *O. sinensis* peaked at Ks *c.* 0.15 (Fig. 2) and from this we estimated divergence occurred around 12 Ma using the mutation rate of $1.2 \times 10^{-8}$ (which lies between 11 and 15 Ma using the lowest, $1 \times 10^{-8}$, and highest, $1.5 \times 10^{-8}$, mutation rates for plants respectively). Similarly,

the Ks values of fourfold degenerate sites for paralogous pairs of sequences was estimated in turn for *O. digyna* and *O. sinensis* (Fig. 2). Though we did not find an obvious peak in the distribution of Ks values for paralogous pairs in *O. digyna*, we found two peaks in the distribution of such Ks values in *O. sinensis* (Fig. 2 and Fig. S3), indicating two possible genome duplications in the ancestry of this species. The first Ks peak at *c.* 0.15 was estimated (Ks/substitution rate) to have occurred around 12.2 Ma (using the synonymous substitution rate $1.2 \times 10^{-8}$, which lies between 11 and 15 Ma using the lowest and highest substitution rates for plants, respectively). This corresponds to the estimated time of divergence between the two species based on the analysis of orthologous pairs. The second genome duplication (a Ks peak at *c.* 0.077) was estimated as above to have occurred 6.0 (5-7) Ma.

The two *Rheum* species and *O. digyna* included in the transcriptome analysis are diploids, while *O. sinensis* is a polyploid in which one or two genome duplications have occurred. We therefore identified 1:1:1:2 and 1:1:1:3 homologous sequences across these four species (using MultiParanoid) with 556 of the first type (1:1:1:2) and 11 of the second kind (1:1:1:3) detected. Two types of gene tree were recovered for the 1:1:1:2 homologous gene datasets (Fig. 3a, b). In one gene tree, two homologous gene sets from *O. sinensis* clustered together (Fig. 3a. 207/556 groups, 37%) while the one set from *O. digyna* comprised a separate clade. In the other tree, one homologous gene set from *O. sinensis* clustered with the only one from *O. digyna* into a monophyletic clade, while the other set from *O. sinensis* comprised an independent clade (Fig. 3b, 349/556 groups, 63%). Similarly, two types of gene tree were obtained for the 1:1:1:3 homologous gene sets (Fig. 3c, d). In one tree, two homologous gene sets from *O. sinensis* occurred together in one monophyletic clade (Fig. 3c. 6/11 groups, 55%) while a third set from *O. sinensis* clustered into a different monophyletic clade with the

only set representing *O. digyna*. In the other tree, the three homologous gene sets from *O. sinensis* formed three separate clades, one of which also contained the set representing *O. digyna* (Fig. 3d, 5/11 groups, 45%).

**Phylogenetic relationships and demographic history based on analyses of seven nuclear loci**

The seven nuclear genes sequenced in *O. sinensis* ranged from 318 to 774 bp in length. The full length of all aligned sequences was 3768 bp. The number of haplotypes recovered for each gene in *O. sinensis* varied from 18 to 24. To assess phylogenetic relationships between haplotypes for each of the seven genes across both *Oxyria* species we constructed haplotype trees for each gene in turn. These trees (Fig. 4) tended to support the topologies of the phylogenetic trees constructed for homologous gene sets identified in the transcriptome analysis. For example, two gene trees (genes 3375 and 6500, Fig. 4) agreed well with the topology of the phylogeny illustrated in Fig. 3d with paralogous alleles of *O. sinensis* comprising two separate clades and orthologs of the two species forming a monophyletic clade. Similarly, the trees obtained for genes 7783 and 1007 are consistent with the phylogenetic trees illustrated in Fig. 3b and Fig. 3c, respectively, while the topologies of the trees for genes 10702 and 9342 agree well with the tree illustrated in Fig. 3a.

ABC modeling and tests of the demographic dynamics of the two *Oxyria* species conducted on population genetic data derived from the seven nuclear genes supported a model of expansion-shrinkage and re-expansion (scenario 4) during the early demographic histories of both species (Fig. 5b). The results of PCA conducted on summary statistics for the four simulated demographic

scenarios, simulations from the posterior predictive distribution, and the observed data showed that the observed data sets for both species were embedded within the many simulated data sets generated for scenario 4 (Fig. S4). Thus, the observed data sets for both species were more similar to the simulated data sets generated for scenario 4 than to those generated for the three other scenarios. For both species, the posterior probabilities obtained for direct estimate and logistic regression were also significantly higher for scenario 4 (direct estimate: P=0.63 to 0.8 for *O. sinensis*, P=0.78 to 1 for *O. digyna*; logistic regression: P=0.60 to 0.70 for *O. sinensis*, P=0.58 to 0.62 for *O. digyna*) than for the other scenarios (Fig. S5). Thus, *O. sinensis* and *O. digyna* are indicated to have experienced two expansions and a bottleneck within the past several million years. The parameter estimates (Table S4) revealed the time of the first expansion (t2) for *O. sinensis* to be 6-7 Ma, which is far earlier than that for *O. digyna* (3-3.2 Ma). Shrinkage in the population sizes of both species is estimated to have occurred between 2.5 and 2 Ma with the time of the second expansion (t1) of *O. sinensis* (db=0.66 Ma, t1=1.38 Ma) being similar to that estimated for *O. digyna* (db=0.74 Ma, t1=1.78 Ma). These latter results therefore suggest that both *O. sinensis* and *O. digyna* experienced similar demographic histories in response to early Pleistocene climate change.

**Ecological niche divergence and distributional shifts in response to recent Quaternary climate change**

Ecological niche modeling (ENM) showed that the environmental variables that contributed most to model predictions were isothermality and precipitation of the warmest quarter for *O. sinensis*, and mean temperature of the wettest quarter and isothermality for *O. digyna*. All models exhibited high predictive ability with AUC>0.92 and SD<0.01. Multiple niche overlap values and identity tests

measured by I and D demonstrated that observed values were significantly higher than actual values (I=0.665, D=0.372) (Fig. 6b), indicating clear ecological differentiation between the two species. ENM further indicated that the distributions of both *Oxyria* species in the parts of Asia investigated in the present study were much narrower during the LIG than at the LGM or currently (Fig. 6a), and that at all three stages the distribution of *O. digyna* was more extensive than *O. sinensis.* The high concordance of predicted areas for both species between the LGM and the present day suggests that little change has occurred in the distributions of these two species over the last 21,000 years.

**Discussion**

The major aim of the research reported here was to reconstruct the evolutionary and demographic history of an allopolyploid species that has one parent species now extinct. We focused on the polyploid *Oxyria sinensis* (2*n*=40), which is one of only two species comprising the genus *Oxyria*, the other being the diploid, *O. digyna* (2*n*=14). GISH established that *O. sinensis* is an allopolyploid containing the entire diploid chromosome complement of *O. digyna* in addition to 26 chromosomes derived from an unknown, and presumably extinct 'ghost' parent species. Reconstructing the ancestry and evolution of such an allopolyploid presents a major challenge, but can proceed through genomic and genetic analyses. These analyses combined with coalescent simulations and ecological niche modeling have also shed light on the historical demography of *O. sinensis* and its extant parent *O. digyna*.

**Origin of *O. sinensis***

From the distribution of Ks values obtained for orthologous sequences in *O. sinensis* and *O. digyna,* identified by transcriptome analysis, moderate signatures of two independent genome duplications were detected in *O. sinensis*, whereas none were detected for *O. digyna* (Fig. 2 and Fig. S3). The first genome duplication was estimated to have occurred around 12 (11-15) Ma. This date is largely consistent with that for divergence of *O. sinensis* from *O. digyna* based on chloroplast DNA sequence variation (Sun *et al.* 2012), which would be unaffected by genome duplication. It is likely, therefore, that the initial genome duplication occurred when the tetraploid ancestor of *O. sinensis* diverged from *O. digyna,* and indeed such duplication may have triggered this divergence. The second genome duplication was estimated to have occurred around 6 (5-7) Ma (Fig. 2). Phylogenetic analyses of homologous sequences detected by transcriptome analysis and also sequences obtained for seven specific nuclear loci, indicated that *O. sinensis* contained the genome of an unknown 'ghost' lineage in addition to a genome similar to that of *O. digyna* (Fig. 3, Fig. 4). This finding, therefore, provided further evidence that *O. sinensis* is an allopolyploid species and that *O. digyna* is one of its progenitors.

*O. sinensis* may have originated in several different ways. First, two unknown and extinct diploid species (both with $2n = 14$ or one with $2n = 12$) could have hybridized and formed an allotetraploid during the first genome duplication approximately 12 Ma, which then hybridized with diploid *O. digyna* to produce hexaploid *O. sinensis* around 6 Ma. Second, diploid *O. digyna* may have hybridized with an unknown extinct diploid species to form a tetraploid around 12 Ma, which then hybridized with a third diploid species (now extinct) to produce hexaploid *O. sinensis* around 6 Ma. Both of these

pathways are supported by those phylogenies constructed from homologous sequences occurring in a 1:3 ratio for *O. digyna* vs. *O. sinensis*, in which one of the three sets present in *O. sinensis* clusters with the set present in *O. digyna,* while the other two sets form two independent clades (Fig. 3d). Third, an unknown diploid species may have undergone genome duplication to produce an autotetraploid around 12 Ma, which later hybridized with *O. digyna* to form *O. sinensis* approximately 6 Ma. This pathway is supported by phylogenies in which one of the homologous gene sets in *O. sinensis* clusters with the set present in *O. digyna,* while the other two sets form a monophyletic clade (Fig. 3c). In all of these cases, if the base chromosome number of the genus is x = 7 (with $2n$ =14 for the extinct diploid species), it may be assumed that two chromosomes were lost after the first or second polyploidization events to produce the chromosome number of $2n$ = 40 for *O. sinensis*.

Of the three pathways described, the second one, involving the formation of an allotetraploid between *O. digyna* and an extinct diploid species as a first step, seems unlikely due to the following. The initial genome duplication (polyploidization) was dated to 12 Ma based on genomic divergence between *O. digyna* and *O. sinensis* according to the distribution of Ks values for 1:1 orthologous sequence pairs (Fig. 2). This date is consistent with the date of divergence between the two species based on maternally inherited chloroplast DNA sequence variation (Sun *et al.* 2012). It is highly unlikely that these two events, i.e. divergence between *O. digyna* and another diploid species, and genome duplication giving rise to an allotetraploid between these two species, would have occurred more or less at the same time. Though we might rule out the second pathway of origin of *O. sinensis*, we are unable to distinguish clearly between the first and third pathways from the available evidence. This is because the phylogenies constructed from homologous sequences supporting one or the other

of these two pathways were recovered (Fig. 3 and 4). Phylogenetic relationships of homologous sequences shared between a polyploid and its ancestors are complicated by DNA loss and concerted evolution, both of which are very common in polyploids (Soltis & Soltis 2016; Alexander-Webber *et al.* 2016). As a consequence, different phylogenies supporting different evolutionary scenarios are likely to be observed across different homologous sequences as was the case here. We can conclude, however, that the origin of a tetraploid (allotetraploid in one case and autotetraploid in the other case) was a first step in the origin of *O. sinensis* (Fig. 7), and that the maternal parent of this tetraploid was a diploid species that diverged from *O. digyna* around 12 Ma.

If *O. sinensis* originated either through the first or third pathway, i.e. involving allopolyploidization 6 Ma between diploid *O. digyna* and an extinct tetraploid 'ghost' species (Fig. 7), it would indicate that though divergence between *O. sinensis* and *O. digyna* can be dated to around 12 Ma, the actual origin of *O. sinensis* is much younger, occurring approximately 6 Ma. This raises an important point regarding dating the origins of allopolyploid species in that depending on their evolutionary history, dates of origin estimated directly from genetic divergence between sister species may be incorrect as could be the case for *O. sinensis*. In many such cases, therefore, detailed analyses of genomic data will be required to reconstruct the evolutionary history of an allopolyploid and determine its true date of origin.

**Demographic history of *O. sinensis* and *O. digyna* in response to past climatic oscillations**

Coalescent simulations using approximate Bayesian computation on population genetic data for seven nuclear genes indicated that *O. sinensis* expanded its population size rapidly approximately 6 Ma following its origin (Fig. 5). The simulations further indicated that towards the end of the Pliocene and beginning of the Pleistocene population sizes of both *O. sinensis* and *O. digyna* contracted markedly, but then expanded again approximately 1.78 Ma (for *O. digyna*) and 1.38 Ma (for *O. sinensis*). Additional and more recent contractions and expansions in the distributions of both species were indicated by ecological niche modeling. ENM indicated that the distributions of both species were much narrower during the Last Interglacial period (0.14-0.12 Ma) than they were at the Last Glacial Maximum (0.021-0.018 Ma) and are at the present time (Fig. 6). Throughout these periods, *O. digyna* maintained a considerably wider distribution than *O. sinensis* in central, south and southwest Asia, and there appears to have been little change in the respective distributions of these species in this region during the last 21,000 years, that is since the LGM.

At some stage during the last 6 million years the tetraploid parent species of *O. sinensis* became extinct with only the presence of its genome in *O. sinensis* signifying its former existence. Similarly, at some stage during the last 12 million years the unknown diploid species involved in the origin of this tetraploid has (or have) also become extinct. It is feasible that *O. sinensis* replaced its tetraploid parent (now extinct) during its rapid expansion of population size that followed its origin approximately 6 Ma. Alternatively, the tetraploid may have become extinct at a later stage during the Pleistocene climatic oscillations that triggered marked fluctuations in species' population sizes. Our current analyses cannot distinguish between these or other possible times when the tetraploid 'ghost'

ancestor became extinct. The results of our analyses do show, however, that such a tetraploid species existed and that its genome currently remains largely intact within the hexaploid *O. sinensis*. It is feasible that this tetraploid became extinct because the combination of its genome with that of *O. digya* in *O. sinensis*, resulted in the origin of a species that was better adapted to meet the challenges of massive climatic changes that occurred in the region over the last 6 million years. What caused the extinction of the ancestral diploid species of *O. sinensis* also cannot be ascertained. This (or these) diploid species may have been replaced by the tetraploid, by *O. sinensis* or *O. digyna*, or became extinct due to other causes, possibly climatic, that occurred in the region during the last 12 million years.

**Acknowledgements**

This work was supported by grants from National Natural Science Foundation of China (31590821), National Key Project for Basic Research (2014CB954100), Ministry of Science and Technology of the People's Republic of China (2010DFA34610) and International Collaboration 111 Projects of China.

**References**

Alexander-Webber D, Abbott RJ, Chapman MA (2016) Morphological convergence between an allopolyploid and one of its parental species correlates with biased gene expression and DNA loss. *Journal of Heredity* 107, 445-454.

Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22, e9-e15.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA (2016) On the relative abundance of autopolyploids and allopolyploids. *New Phytologist*, 210, 391-398.

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *Plant Cell*, 16, 1667-78.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

Brochmann C, Brysting AK, Alsos IG *et al.* (2004) Polyploidy in arctic plants. *Biological Journal of the Linnean Society*, 82, 521-536.

Cornuet J, Ravigne´ V, Estoup A (2010) Inference on population history and model checking using DNA sequence and microstaellite data with the software DIYABC (v.1). *BMC Bioinformatics*, 11.

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem*, Bull 19, 11-15.

Doyle JJ, Sherman-Broyles (2016) Double trouble: taxonomy and definitions of polyploidy. *New Phytologist* (in press) DOI: 10.1111/nph.14276

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. *Proceeding of National Academy of Sciences*, 106, 5737-5742.

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24, 38-49.

Fowler NL, Levin DA (1984) Ecological constraints on the establishment of a novel polyploid in competition with its diploid progenitor. *The American Naturalist*, 124, 703–711.

Fraley C, Raftery AE, Murphy TB, Scrucca L (2012) mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29, 644-652.

Grant V (1981) *Plant speciation*. Columbia University Press, New York.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59, 307-321.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25, 1965-1978.

Holliday JA, Zhou L, Bawa R, Zhang M, Oubida RW (2016) Evidence for extensive parallelism but divergent genomic architecture of adaptation along altitudinal and latitudinal gradients in *Populus trichocarpa*. *New Phytologist*, 209, 1240-1251.

Huang Y, Niu B, Gao Y, Fu L, Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26, 680-682.

Jewell DC, Islam-Faridi N (1994) A technique for somatic chromosome preparation and C-banding of maize. In *the Maize Handbook*, Inc. pp. 484-493. Springer press, New York.

Jiao Y, Leebens-Mack J, Ayyampalayam S *et al.* (2012) A genome triplication associated with early diversification of the core eudicots. *Genome Biology*, 13, 1.

Leitch AR, Schwarzacher T, Jackson D, Leitch IJ (1994) In situ hybridization: a practical guide. *BIOS Scientific Publishers Ltd*.

Lewis WH (1980) Polyploidy in angiosperms: dicotyledons. *In Polyploidy*, pp, 241-268. Springer press, New York.

Liu FH, Yu FH, Liu WS *et al.* (2007) Large clones on cliff faces: expanding by rhizomes through crevices. *Annals of Botany*, 100, 51-54.

Liu JQ (2004) Uniformity of karyotypes in *Ligularia* (Asteraceae: Senecioneae), a highly diversified genus of the eastern Qinghai-Tibet Plateau highlands and adjacent areas. *Botanical Journal of the Linnean Society*, 144, 329-342.

Liu JQ, Ho TN, Liu SW, Lu AM (2001) Karyological studies on the Sino-Himalayan endemic genus, *Cremanthodium* (Asteraceae: Senecioneae). *Botanical Journal of the Linnean Society*, 135, 107-112.

Löytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320, 1632-1635.

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science*, 290, 1151-1155.

Meng L, Chen G, Li Z, Yang Y, Wang Z, Wang L (2015) Refugial isolation and range expansions drive the genetic structure of *Oxyria sinensis* (Polygonaceae) in the Himalaya-Hengduan Mountains. *Scientific reports*, 5.

Mittermeier R *et al.* (2004) *Hotspots revisited: Earth's biologically richest and most endangered terrestrial ecoregions*. CEMEX & Agrupacion Sierra Madre, Mexico.

Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GA, Kent J (2000) Biodiversity hotspots for conservation priorities. *Nature*, 403, 853-858.

Nie ZL, Wen J, Gu ZJ, Boufford DE, Sun H (2005) Polyploidy in the flora of the Hengduan Mountains hotspot, southwestern China. *Annals of the Missouri Botanical Garden*, 92, 275-306.

Ossowski S, Schneeberger K, Lucas-Lledó JI *et al.* (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, 327, 92-94.

Otto SP (2007) The evolutionary consequences of polyploidy. *Cell*, 131, 452-462.

Parisod C, Broennimann O (2016) Towards unified hypotheses of the impact of polyploidy on ecological niches. *New Phytologist*, 212, 540-542.

Phillips SJ, Dudík M (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, 31, 161-175.

Ramsey J (2011) Polyploidy and ecological adaptation in wild yarrow. *Proceedings of the National Academy of Sciences*, 108, 7096-7101.

Ramsey J, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics*, 29, 467–501.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference nder mixed models. *Bioinformatics*, 19, 1572-1574.

Scott LS, Stenz NW, Ingvarsson P, Baum DA (2016) Whole genome duplication in coast redwood (*Sequoia sempervirens*) and its implications for explaining the rarity of polyploidy in conifers. *New Phytologist*, 211, 186-193.

Soltis PS, Soltis DE (2016) Ancient WGD events as drivers of key innovations in angiosperms. *Current Opinion in Plant Biology*, 30, 159-165.

Soltis DE, Visger CJ, Soltis PS (2014) The polyploidy revolution then…and now: Stebbins revisted. *American Journal of Botany*, 101, 1057-1078.

Sonnhammer EL, Östlund G (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, 43, D234-D239.

Stebbins GL (1971) *Chromosomal evolution in higher plants.* Addison-Wesley, London.

Stebbins GL (1985) Polyploidy, hybridization, and the invasion of new habitats. *Annals of the Missouri Botanical Garden*, 72, 824-832.

Sun YS, Wang AL, Wan DS, Wang Q, Liu JQ (2012) Rapid radiation of *Rheum* (Polygonaceae) and parallel evolution of morphological traits. *Molecular Phylogenetics and Evolution*, 63, 150-158.

Swets JA (1988) Measuring the accuracy of diagnostic systems. *Science*, 240, 1285–1293.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, 30, 2725-2729.

Wang L, Zhou H, Han J, Milne RI, Wang M, Liu B (2014) Genome-Scale Transcriptome Analysis of the Alpine "Glasshouse" Plant Rheum nobile (Polygonaceae) with Special Translucent Bracts. *PLoS ONE*, 9, e110712.

Wang Q, Liu J, Allen GA *et al.* (2016) Arctic plant origins and early formation of circumarctic distributions: a case study of the mountain sorrel, *Oxyria digyna*. *New Phytologist*, 209, 343-353.

Warren DL, Glor RE, Turelli M (2010) ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, 33, 607-611.

Wilson EO (1992) *The diversity of life*. Harvard University Press, Boston.

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon, PB, Rieseberg LH (2009) The frequency of polyploid speciation in vascular plants. *Proceeding of National Academy of Sciences*, 106, 13875-13879.

Wu Z (1998) *Delineation and unique features of the Sino-Japanese floristic region*. University of Tokyo Bulletin, Tokyo.

Yang J, Wang ZK, Zhu WL, Meng LH (2014) Responses to drought stress among sex morphs of *Oxyria sinensis* (Polygonaceae), a subdioecious perennial herb native East Himalayas. *Ecology and Evolution*, 4, 4033-4040.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586-1591.

Zhao F, Yang YP (2008) Reproductive allocation in a dioecious perennial *Oxyria sinensis* (Polygonaceae) along altitudinal gradients. *Journal of Systematics and Evolution*, 46, 830-835.

**Author contributions**

The authors collaborate on projects on the origin and evolution of the species in China. J.L. designed the research; X.L., Q.H., X.Z., D.Z. and Q.W. performed the research; X.L. and Q.H analyzed the data; J.L. and R.A. wrote the manuscript.

**Data accessibility**

NCBI SRA: SRX621187, SRX 2200537, SRX2200544, SRX2200877.

Accession number: KX999721-KX999862, KR003454-KR003705.

The locations of samples used in ENM analysis are archived in a directory with the name 'ENM_data', which can be downloaded from Dryad doi:10.5061/dryad.bc1v8.

**Supporting information**

Additional supporting information may be found in the online version of this article.

**Fig. S1** Global distributions of *O. sinensis* (in blue) and *O. digyna* (in green). Migration routes of *O. digyna* from the QTP into the arctic region and beyond are indicated by arrows (from Wang *et al*. 2016). Abbreviations represent Qinghai-Tibet Plateau (QTP), Russia (RU), Europe (EU), North America (NA) and Greenland (GR) respectively.

**Fig. S2** Geographical distribution of *O. sinensis* and *O. digyna* samples used in this study (with populations 19 and 20 excluded). Detailed information on populations is provided in Table S1

**Fig. S3** Ks value frequency for fourfold degenerate sites of *O. sinensis.* Evidence of first duplication is colored green, Ks = 0.153. Evidence of secondary duplication is colored blue, Ks = 0.076. Normal mixture models (black curve) were estimated by mclust.

**Fig. S4** Pre-evaluate scenario-prior combinations for (a) *O. sinensis* and (b) *O. digyna* in QTP. Model checking results showing the first 2 axes of a PCA on summary statistics from the 4 simulated demographic scenarios, simulations from the posterior predictive distribution, and the observed data.

**Fig. S5** Posterior probabilities for the 4 competing demographic scenarios for (a) *O. sinensis* and (b) *O. digyna* in the QTP from both a direct estimate using the 500 closest datasets and a logistic regression using 1% of closest datasets.

**Table S1** Sampling localities of *O. sinensis* and *O. digyna* used in this study. Sampling localities (except for populations 19 and 20) are shown in Figure S2.

**Table S2** Assembly statistics for the two *Oxyria* species.

**Table S3** Descriptions and prior settings for all parameters used in DIYABC (four scenarios for all *O. sinensis* and four scenarios for *O. digyna* in the QTP). Population size parameters are in units of

population effective size (N), while time parameters (including bottleneck duration) are in units of generations.

**Table S4** Estimations of the posterior distributions of parameters revealed by Approximate Bayesian Computation for the best scenario of the demographic history of *O. sinensis* and *O. digyna* in the QTP. Estimation is based on 1% of the closest simulated data sets and the logit transformation of parameters was used. Time parameter (t and db) (unit of time: years) was scaled by generation time (5 years).

**Fig. 1** Chromosomes of (a) *O. sinensis* and (b) *O. digyna* at mitotic metaphase. (c) *Oxyria sinensis* (2*n*=40) chromosomes at mitotic metaphase probed with labeled total genomic DNA of *O. digyna* (2*n*=14). GISH shows 14 chromosomes (green) of *O. sinensis* containing DNA similar to that of *O. digyna*. GISH is not evident for the remaining 26 chromosomes of *O. sinensis* (stained blue) showing that these are not derived from *O. digyna*. Bar = 9 um.

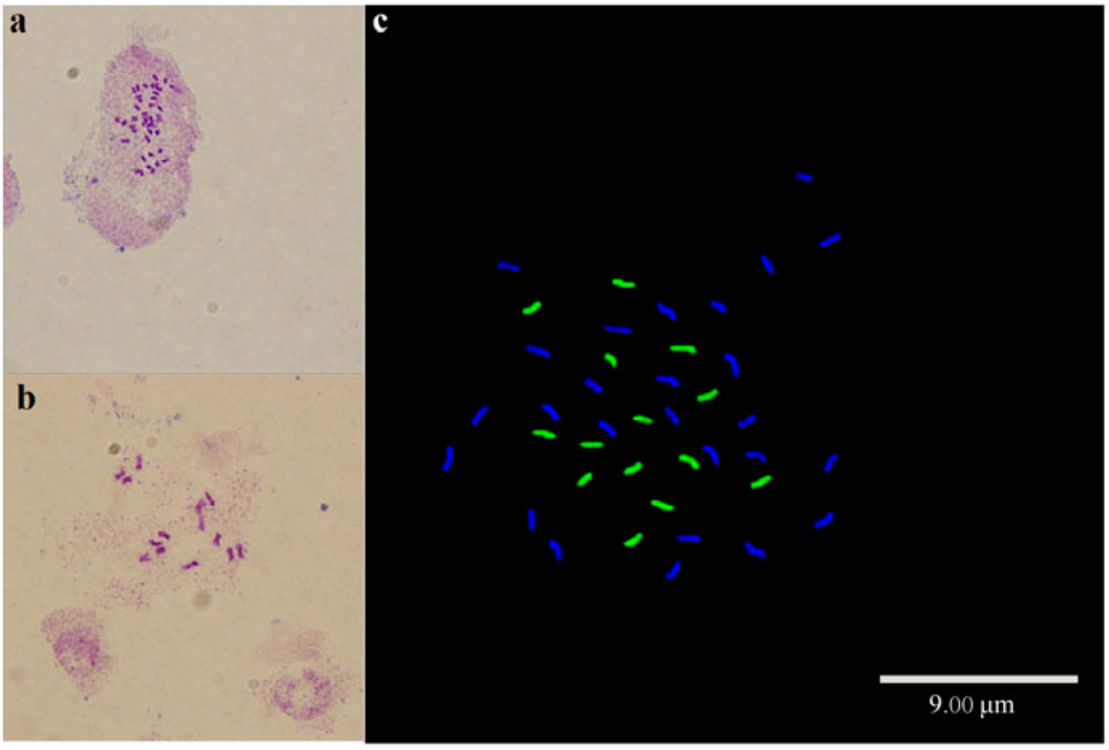

**Fig. 2** Ks value distribution of fourfold degenerate sites of orthologs (yellow line) and paralogs (blue and green lines) between and within two *Oxyria* species to identify and estimate (Ks/mutation rate) interspecific divergence and genome duplications. d4= The Ks value of fourfold degenerate sites, density=The frequency of d4.

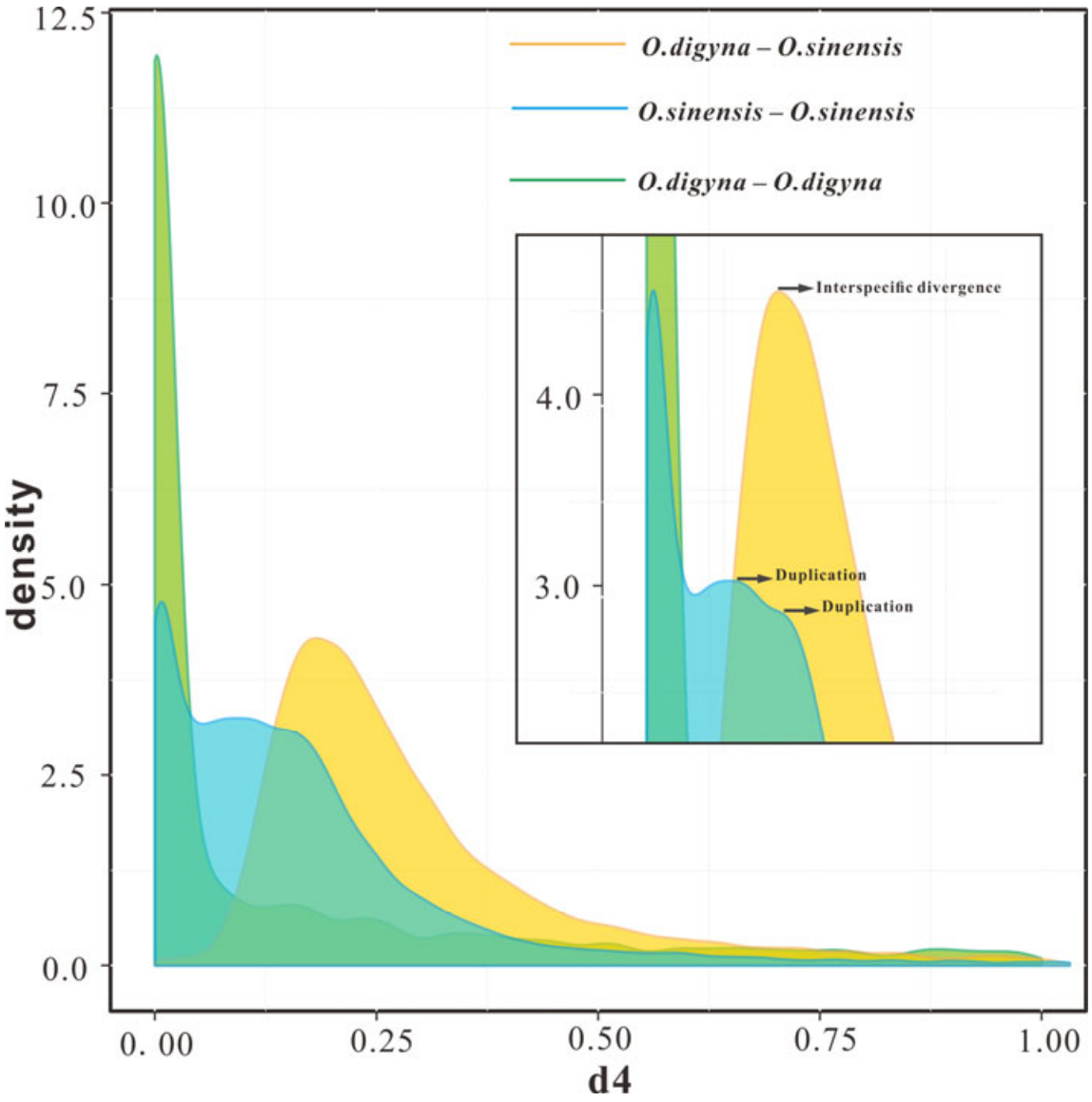

**Fig. 3** The homologous gene relationships from maximum likelihood analysis based on homologous sequence identified by MultiParanoid according to ratios of 1:1:1:2 (a and b) and 1:1:1:3 (c and d). The number and percentage of gene groups are presented for each tree type. Species are illustrated by blue (*O. sinensis*), green (*O. digyna*), yellow (*Rh. tanguticum*) and orange (*Rh. nobile*) respectively.

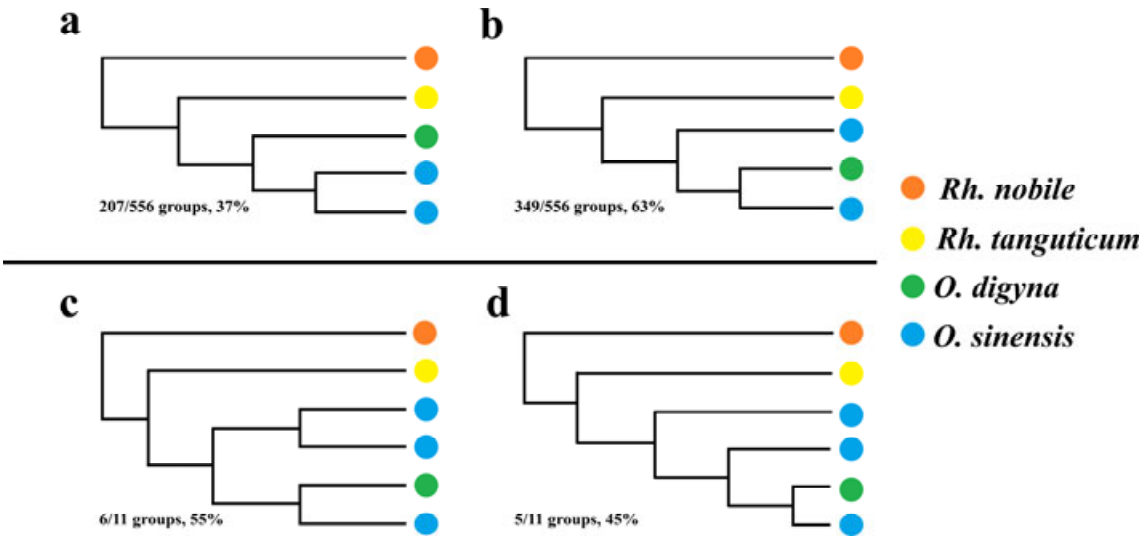

**Fig. 4** Bayesian phylogenetic trees constructed for haplotypes of seven nuclear loci present in *O. digyna* and *O. sinensis*, with clade support indicated by Bayesian posterior probabilities. *Rh. tanguticum* (triangle) was used as outgroup.

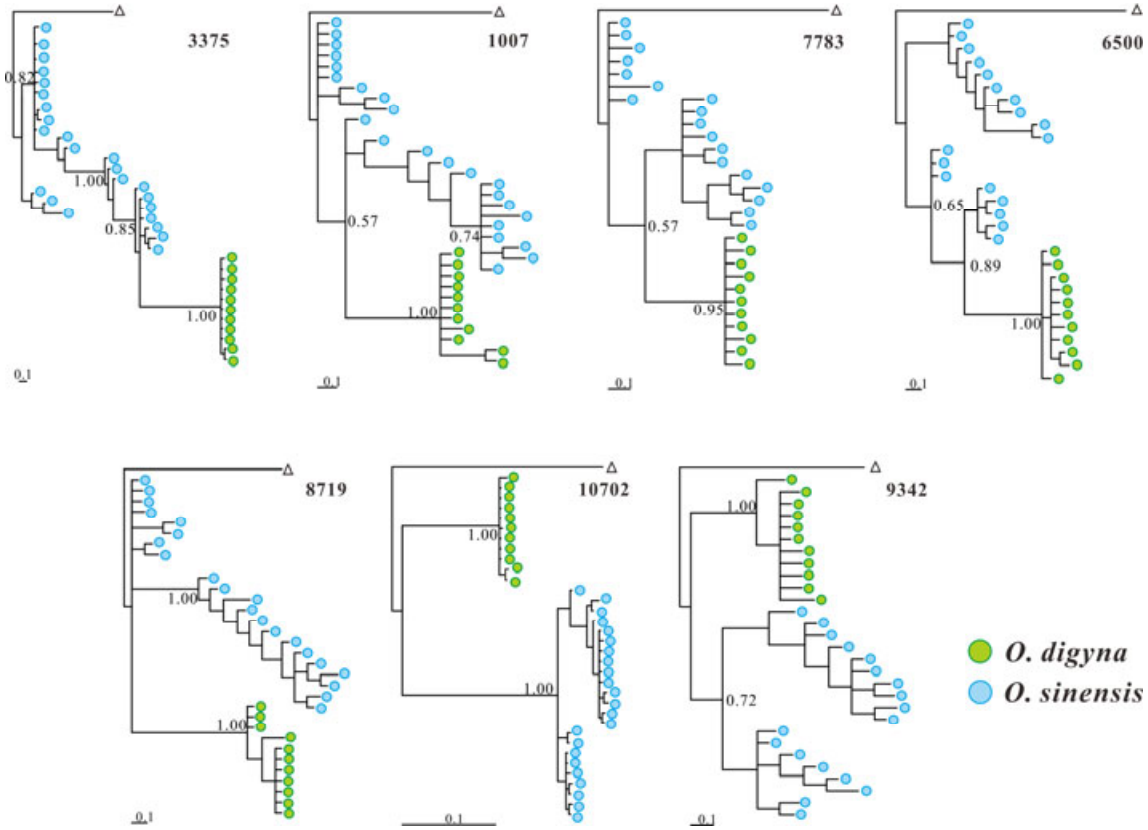

**Fig. 5** (a) Schematic diagram of the four demographic scenarios tested by DIYABC. (b) Inferred demographic history scenarios for the two *Oxyria* species. NA, ancestral population size; N1, current population size; N1b, contracted or expanded population size; t, times in years from present; db, duration of bottleneck; Ma, million years ago.

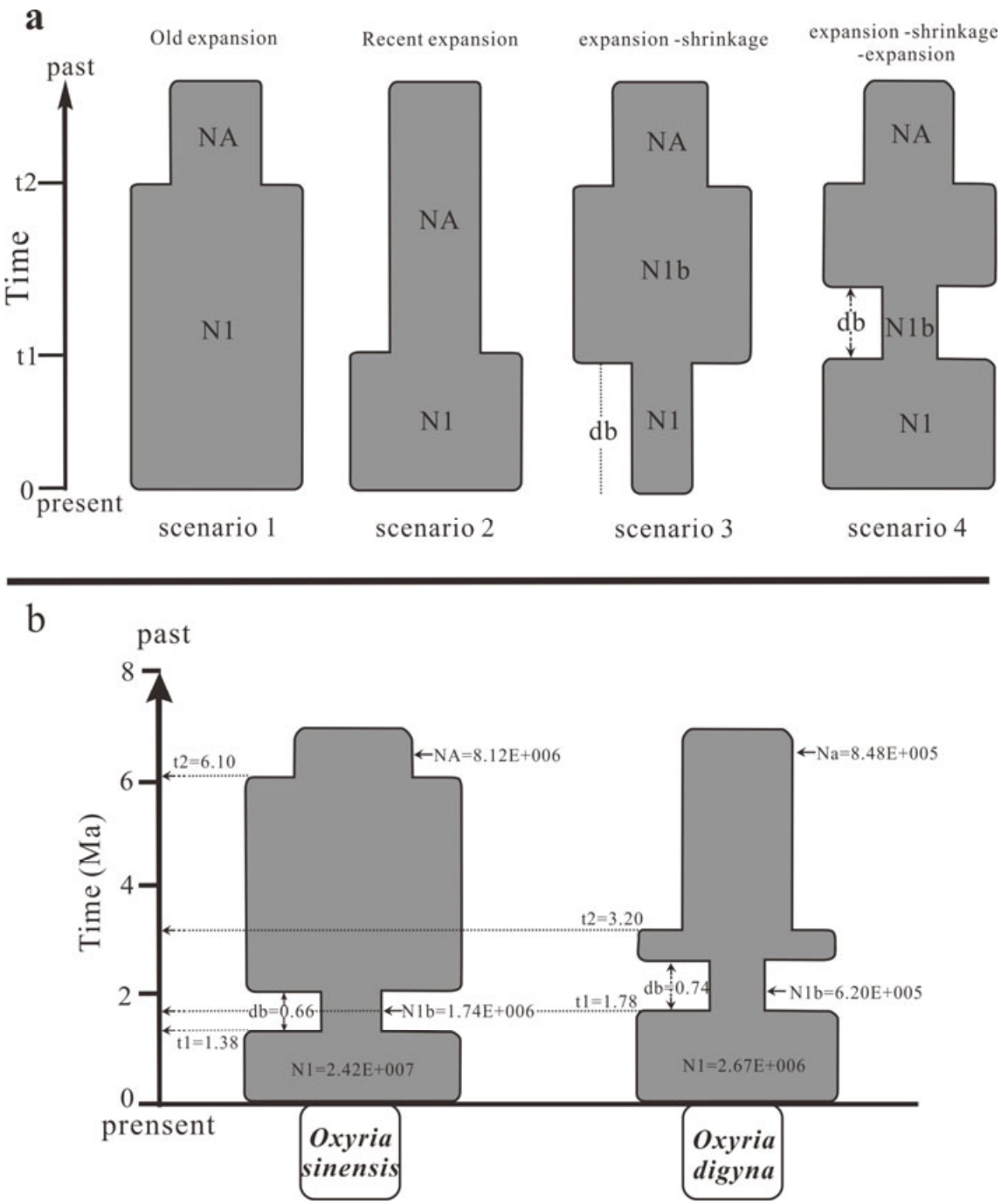

**Fig. 6** Ecological niche modeling of differentiation between *O. sinensis* and *O. digyna*. (a) Predicted distributions of *O. sinensis* and *O. digyna* in central, south and south-west Asia at the present time, the LGM and LIG. MYA, million years ago. Colours represent the probability of a species occurring in an area, from highest (red) to lowest (gray). (b) The results of niche identity tests (I and D) between *O. sinensis* and *O. digyna*. Gray bars show the simulated identity values generated from 100 randomizations, arrows indicate actual values from MaxEnt modeling.

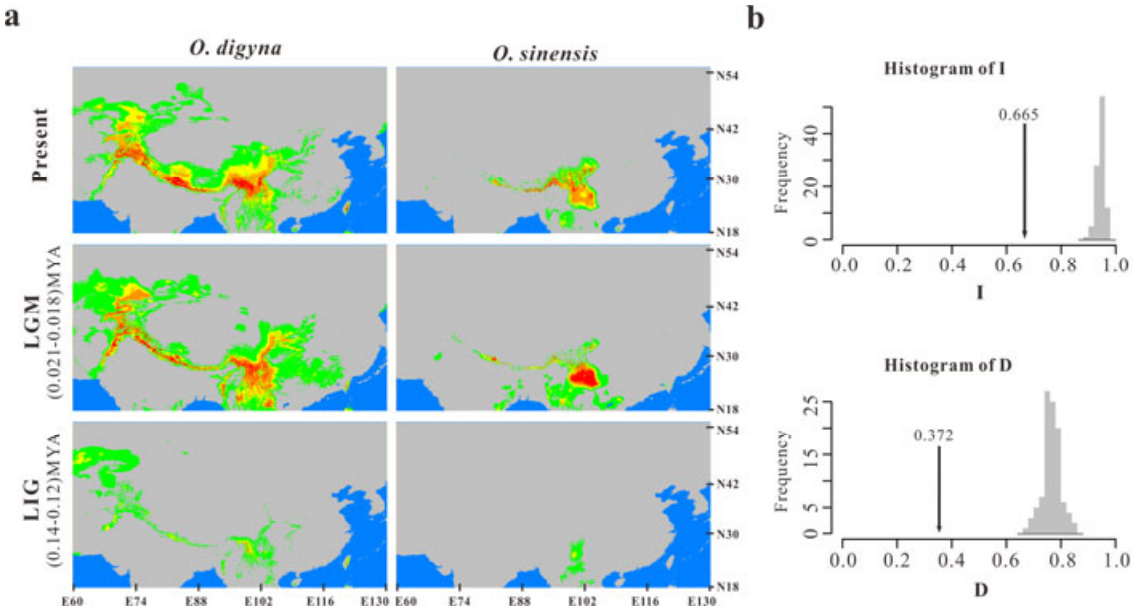

**Fig. 7** A diagram portraying the allopolyploid origin of *O. sinensis* between *O. digyna* and the unknown 'ghost' species. Times (horizontal axis) in grams are assumed to be proportional. Ma, million years ago. The solid lines indicate extant species while dotted lines indicate the unknown or extinct species. Number of lines are proportional to the ploidy of the species.